PROD. TYPE: COM
PP: 1-13 (col.fig.: nil)

JSPI 3218 MODC+

ED: Bhagya
PAGN: Suresh – SCAN: Global

ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

Journal of Statistical Planning and Inference ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Log-linear models for mutations in the HIV genome☆

C. Ahn[a], G.G. Koch[a,*], L. Paynter[a], J.S. Preisser[a], F. Seillier-Moiseiwitsch[b]

[a]*Department of Biostatistics, School of Public Health, Campus Box 7420, Chapel Hill, NC 27599-7420, USA*
[b]*Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington, DC 20057-1484, USA*

Abstract

We discuss a general application of categorical data analysis to mutations along the HIV genome. We consider a multidimensional table for several positions at the same time. Due to the complexity of the multidimensional table, we may collapse it by pooling some categories. However, the association between the remaining variables may not be the same as before collapsing. We discuss the collapsibility of tables and the change in the meaning of parameters after collapsing categories. We also address this problem with a log-linear model. We present a parameterization with the consensus output as the reference cell as is appropriate to explain genomic mutations in HIV. We also consider five null hypotheses and some classical methods to address them. We illustrate methods for six positions along the HIV genome, through consideration of all triples of positions.
© 2007 Published by Elsevier B.V.

MSC: 62F03; 62H17; 62H20; 62J12; 62P10

Keywords: HIV genome; Consensus; Correlated mutation; Collapsibility; Conditional association; Marginal association; Log-linear models

1. Introduction

Nucleotides and amino acids can be viewed as categorical data. In this paper, we are interested in mutations along the HIV genome. Like other DNA or RNA sequences, HIV sequences are coded by means of four nucleotides or 20 amino acids. These are just nominal variables. We consider amino-acid sequences, since the information contained in amino acids is more important from a functional/structural point of view.

Mutations are usually deleterious to any organism. However, in the case of HIV, a mutation can provide a virus with a better chance to escape the host immune system, which helps the virus survive and pass on its genetic material. We believe mutations at specific positions are correlated, and tendencies for some combinations of mutations to be observed more frequently than expected by chance provide proof. These double mutations either maintain the structure of a vital protein (possibly when a single mutation would destabilize this structure) or yield a viable structural form not recognized by the host.

Several positions along the HIV genome are simultaneously represented by a multidimensional contingency table. High-dimensional tables can have complicated structures and interpretations for model parameters. Roy and Mitra (1956), Roy and Kastenbaum (1956), Roy (1957), Roy and Bhapkar (1960) and Bhapkar and Koch (1968) describe

☆ This research was partially supported by Grant R01 AI47068 from the National Institutes of Health.
* Corresponding author.
E-mail address: bcl@bios.unc.edu (G.G. Koch).

0378-3758/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.jspi.2007.03.007

Please cite this article as: Ahn, C., et al., Log-linear models for mutations in the HIV genome, J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.007

UNCORRECTED PROOF

the structure and several hypotheses of interest for three-way tables. Agresti (2002) and Imrey (2002) describe those hypotheses in terms of a hierarchical log-linear model. Other related discussion is provided in Imrey et al. (1981, 1982, 1996), Imrey (2000) and Imrey and Koch (2005).

Following S.N. Roy's nomenclature, we treat positions as "variates" and do not condition on marginal totals. This approach is motivated by the random nature of the substitution process. In Roy and Kastenbaum (1956) and Roy and Mitra (1956), hypotheses relevant to this set-up were formulated. We will consider, among these, those that are interpretable in our context, namely the hypotheses of conditional and multiple independence (H_{03} and H_{04}, respectively in Section 2.2).

A multidimensional contingency table may contain many zeroes or very small cell counts, and this sparseness of data may undermine test statistics with distributional properties based on large-sample sizes. To use such test statistics, we may have to pool some categories to ensure cell counts are adequate for a large-sample approximation to be applicable. However, the meaning of the parameters may not remain the same. Nevertheless, it will be the same in cases where some collapsibility conditions are satisfied.

In the subsequent sections of this paper, we use the reference-cell parameterization for multidimensional contingency tables since it is appropriate to explain mutations in the HIV genome. We initially consider a model for HIV mutations resulting in a 2×2 table, and then extend it to the $r \times c$ and $2 \times 2 \times 2$ situations. Instead of presenting a $I \times J \times K$ contingency table, we describe it with a log-linear model for simplicity. We investigate the collapsibility of a three-way table by considering its conditional and marginal associations. We identify five hypotheses which might be appropriate for HIV mutations. One of these hypotheses corresponds to an interpretable non-hierarchical model.

For hypothesis testing, we simply use a Wald statistic and/or a deviance test statistic. We illustrate methods with results from an analysis involving six positions. We consider all triples of positions, and we present some of them as examples to explain the five hypotheses of interest.

1.1. Parameterization and hypotheses

We consider models for multinomially distributed cell counts rather than cell probabilities. Based on the biological meaning of mutations and the concept of consensus, we present a reference-cell coding for the model with the consensus as the reference cell.

The usual issue of interest in a contingency table is the independence among two or more categorical variates. However, for our consensus-referencing parameterization, the main objective is the investigation of whether double mutations provide a survival advantage to a virus. Such structure will be considered for the 2×2 table and then extended to the more general $r \times c$ table and the $2 \times 2 \times 2$ table as the simplest three-way table.

1.2. Modelling HIV mutations: the 2×2 case

Fig. 1 shows the parameterization which will be used throughout this paper. For simplicity, we consider two positions with two categories at each position: A_1 and A_2 at the first position and B_1 and B_2 at the second position. If (A_1, B_1) is the original sequence, there are four possible progenies, (A_1,B_1), (A_1,B_2), (A_2,B_1) and (A_2,B_2). (A_1,B_2) has a mutation at the second position, (A_2,B_1) has a mutation at the first position, and (A_2,B_2) has mutations at both positions. We assume that the mutation at the first position is independent of that at the second position, since mutations in HIV are

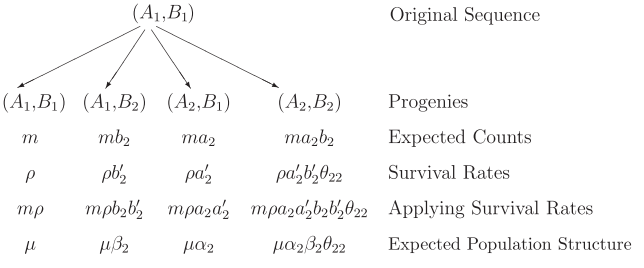

Fig. 1. Parameterization.

UNCORRECTED PROOF

Please cite this article as: Ahn, C., et al., Log-linear models for mutations in the HIV genome, J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.007

Table 1
2 × 2 Table

		Position II	
		B_1	B_2
Position I	A_1	n_{11} μ	n_{12} $\mu\beta_2$
	A_2	n_{21} $\mu\alpha_2$	n_{22} $\mu\alpha_2\beta_2\theta_{22}$

due in the most part to the high error rate of the enzyme reverse transcriptase. In other words, mutations can be thought to happen at random.

Therefore, we apply an independence model to obtain the expected counts for the four pairs, taking (A_1,B_1) as the reference cell: m for (A_1,B_1), mb_2 for (A_1,B_2), ma_2 for (A_2,B_1) and ma_2b_2 for (A_2,B_2). Here, a_2 represents the mutation rate from A_1 to A_2 and b_2 the mutation rate from B_1 to B_2. Since we assumed mutations at different positions are independent, the expected cell count of (A_2,B_2) is ma_2b_2. However, these are the expected numbers of viruses just budding from the infected CD4 cell. Their genetic variation helps in evading the immune system of the host and in successfully generating numerous progenies. But, some mutations are so deleterious that the resulting virions cannot generate vital proteins and disappear.

Thus, we also specify a model for the survival rates. The parameter a_2' represents the single-mutation survival effect, which is the survival rate of mutants with A_2 relative to the strains with A_1. Similarly, b_2' denotes the single-mutation survival effect for B_2. Finally, θ_{22} denotes the double-mutation effect, and measures the surplus effect beyond the multiplicative effects of two combined single mutations.

We compute the expected counts of the surviving viruses by multiplying the expected counts and the corresponding survival rates. After defining $a_2a_2' = \alpha_2$ and $b_2b_2' = \beta_2$, the parameterization in Table 1 is obtained. Note that α_2 and β_2 combine mutation rates (a_2 and b_2) and single-mutation survival effects (a_2' and b_2') since they cannot be estimated separately. However, we can determine the double-mutation survival effect θ_{22} via an odds ratio, since

$$\frac{\mu \times \mu\alpha_2\beta_2\theta_{22}}{\mu\alpha_2 \times \mu\beta_2} = \theta_{22}.$$

The hypotheses of interest for θ_{22} are as follows:

H_0: Double mutation is neutral (i.e. $\theta_{22} = 1$), and
H_a: Double mutation is not neutral (i.e. $\theta_{22} \neq 1$).

1.3. Consensus

Our parameterization is based on mutations, and a mutation is a relatively rare event, even for HIV when considering a specific genomic segment. Let n_{11} denote the cell count for sequences with no mutation at both positions. It is expected to be dominant relative to the other cell counts. We refer to it as the *consensus*. We make use of this information to increase the relevance of a test for the hypothesis H_0. Actually, we use the consensus as a reference and consider the other categories as departures from the consensus. If we use reference-cell coding for a contingency table, the parameterization depends on the choice of cell as the reference. In this case, the consensus is a well-identified cell. Therefore, unlike a usual contingency table, where the reference cell is arbitrary, here the reference cell is well defined.

The objective is not just to test independence between positions, but to evaluate whether double mutations from the consensus affect the survival of the virus.

1.4. The r × c case

If there are r characters at position I and c at position II, the parameterization for the 2×2 case can be easily extended to the $r \times c$ case (Table 2). In this case, some cells with small counts are inevitable, and this sparseness

UNCORRECTED PROOF

Please cite this article as: Ahn, C., et al., Log-linear models for mutations in the HIV genome, J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.007

Table 2
$r \times c$ Table

		Position II			
		B_1	B_2	$\cdots$	B_c
Position I	A_1	$n_{11},\ \mu$	$n_{12},\ \mu\beta_2$	$\cdots$	$n_{1c},\ \mu\beta_c$
	A_2	$n_{21},\ \mu\alpha_2$	$n_{22},\ \mu\alpha_2\beta_2\theta_{22}$	$\cdots$	$n_{2c},\ \mu\alpha_2\beta_c\theta_{2c}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	A_r	$n_{r1},\ \mu\alpha_r$	$n_{r2},\ \mu\alpha_r\beta_2\theta_{r2}$	$\cdots$	$n_{rc},\ \mu\alpha_r\beta_c\theta_{rc}$

Table 3
$2 \times 2 \times 2$ Table

Position I		A_1		A_2	
Position II		B_1	B_2	B_1	B_2
Position III	C_1	n_{111}	n_{121}	n_{211}	n_{221}
		μ	$\mu\beta_2$	$\mu\alpha_2$	$\mu\alpha_2\beta_2\theta_{221}$
	C_2	n_{112}	n_{122}	n_{212}	n_{222}
		$\mu\gamma_2$	$\mu\beta_2\gamma_2\theta_{122}$	$\mu\alpha_2\gamma_2\theta_{212}$	$\mu\alpha_2\beta_2\gamma_2\theta_{122}\theta_{212}\theta_{221}\delta_{222}$

undermines the use of test statistics based on large-sample approximations. For this reason, it is tempting to pool some cells by assuming the θ_{ij}'s in those cells are the same, perhaps because those sparse cells might not have much effect on the test statistic. An extreme version of pooling is the 2×2 case with both positions dichotomous, i.e. consensus vs. non-consensus. An other possibility is to consider a specific row and column together with the consensus row and column as a 2×2 table. This instance is reasonable, when we are interested in a specific pair of mutations. However, the cell sizes in the resulting 2×2 table relative to the consensus should be sufficient to provide adequate power to evaluate this specific pair of mutations.

1.5. *The* $2 \times 2 \times 2$ *case*

For three loci, positions I, II and III, let (A_1, B_1, C_1) denote the predominant configuration so that the number of sequences exhibiting the consensus triplet appears in the (1,1,1)-cell (Table 3). By recognizing n_{111} as the consensus count, we have n_{112}, n_{121} and n_{211} as counts of sequences showing a single mutation from this consensus; also, n_{122}, n_{212} and n_{221} are the counts of sequences with double mutations, and n_{222} the count of sequences with triple mutations. The expected number in the consensus reference cell is μ, while α_2, β_2 and γ_2 are parameters associated with single mutations, θ_{ijk}'s are parameters for double mutations and δ_{222} is the parameter for the triple mutation. The consensus cell ($i = 1$, $j = 1$ and $k = 1$) is used as reference. The expected count of the (2,2,2)-cell involves all the double-mutation and triple-mutation effects.

2. Methods

The $2 \times 2 \times 2$ case, which is the simplest three-way table, has a complex structure. We need to be careful in interpreting the meaning of the parameters: different mechanisms can bring about the observed results.

The 2×2 table for positions I and II can be interpreted as a collapsed $2 \times 2 \times 2$ table by disregarding position III. However, the double-mutation effect θ_{221} has the same effect as in the 2×2 table, only if positions II and III are conditionally independent given position I ($\theta_{122} = \delta_{222} = 1$) or positions I and III are conditionally independent given position II ($\theta_{212} = \delta_{222} = 1$) as shown in Section 2.1. Therefore, investigating double-mutation effects along with the triple-mutation effect is critical to avoid misunderstanding of the complicated relationships.

UNCORRECTED PROOF

Please cite this article as: Ahn, C., et al., Log-linear models for mutations in the HIV genome, J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.007

Table 4
The table after combining position III, where $\mu^* \equiv \mu(1+\gamma_2)$

		Position II	
		B_1	B_2
Position I	A_1	$n_{11\cdot}$ $\mu(1+\gamma_2)$	$n_{12\cdot}$ $\mu\beta_2(1+\gamma_2\theta_{122})$
	A_2	$n_{21\cdot}$ $\mu\alpha_2(1+\gamma_2\theta_{212})$	$n_{22\cdot}$ $\mu\alpha_2\beta_2\theta_{221}(1+\gamma_2\theta_{122}\theta_{212}\delta_{222})$
		⇓	
Position I	A_1	$n_{11\cdot}$ μ^*	$n_{12\cdot}$ $\mu^*\beta_2\left(\frac{1+\gamma_2\theta_{122}}{1+\gamma_2}\right)$
	A_2	$n_{21\cdot}$ $\mu^*\alpha_2\left(\frac{1+\gamma_2\theta_{212}}{1+\gamma_2}\right)$	$n_{22\cdot}$ $\mu^*\alpha_2\beta_2\,\theta_{221}\left(\frac{1+\gamma_2\theta_{122}\theta_{212}\delta_{222}}{1+\gamma_2}\right)$

2.1. Collapsibility of variates

As noted previously, two-way tables can be obtained by collapsing one of the three factors in a three-way table. Table 4 shows the parameterization when position III is collapsed. The interpretation of parameters in the collapsed two-way table may be different from that in the three-way table.

The odds ratio for position I vs. position II in the collapsed table is

$$\frac{(1+\gamma_2)(1+\gamma_2\theta_{122}\theta_{212}\delta_{222})\,\theta_{221}}{(1+\gamma_2\theta_{122})(1+\gamma_2\theta_{212})}.$$

If $(1+\gamma_2)(1+\gamma_2\theta_{122}\theta_{212}\delta_{222})/(1+\gamma_2\theta_{122})(1+\gamma_2\theta_{212})=1$, θ_{221} can be considered as representing the marginal association between positions I and II. Indeed, as recognized in Agresti (1990) and Roy (1957), the following statements are mathematically equivalent:

$$\begin{aligned}&\frac{(1+\gamma_2)(1+\gamma_2\theta_{122}\theta_{212}\delta_{222})}{(1+\gamma_2\theta_{122})(1+\gamma_2\theta_{212})}=1\\ &\quad\Leftrightarrow\quad (1-\theta_{122})(1-\theta_{212})+\theta_{122}\theta_{212}(\delta_{222}-1)(1+\gamma_2)=0\\ &\quad\Leftrightarrow\ \delta_{222}=1\quad\text{and}\quad\{\theta_{122}=1\text{ or }\theta_{212}=1\}.\end{aligned}$$

Therefore, if positions II and III are conditionally independent given position I or if positions I and III are conditionally independent given position II, then θ_{221} represents the marginal association. More generally, it represents the conditional association at the consensus for position III; in addition if $\delta_{222}=1$, it represents the conditional association between positions I and II for both categories of position III.

We introduced the triple-mutation effect δ_{222} in the three-way table. Therefore, we need to pay attention to its role in the two-way table. When there is no double-mutation effect (i.e. $\theta_{122}=\theta_{212}=\theta_{221}=1$), the odds ratio in the marginal table for position I vs. position II is

$$\frac{(1+\gamma_2)(1+\gamma_2\theta_{122}\theta_{212}\delta_{222})\theta_{221}}{(1+\gamma_2\theta_{122})(1+\gamma_2\theta_{212})}=\frac{1+\gamma_2\delta_{222}}{1+\gamma_2}\equiv\varphi,$$

which is not equal to 1 unless $\delta_{222}=1$. This structure implies that the triple-mutation effect can be misinterpreted as a double-mutation effect, when we collapse the three-way table into a two-way table. However, when the triple-mutation

UNCORRECTED PROOF

Please cite this article as: Ahn, C., et al., Log-linear models for mutations in the HIV genome, J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.007

effect is negative ($0 < \delta_{222} < 1$), φ is bounded to a small interval:

$$\frac{1}{1+\gamma_2} < \varphi < 1.$$

The lower bound $1/(1+\gamma_2)$ is obtained when $\delta_{222}=0$, which means that the triple-mutation effect is extremely negative. Since n_{111} is the consensus, γ_2 must be much less than 1; usually, it is rarely over 0.5. When there is a extremely negative triple-mutation effect ($\delta_{222} = 0$), $\varphi = 0.67$ for $\gamma_2 = 0.5$, $\varphi = 0.71$ for $\gamma_2 = 0.2$ and $\varphi = 0.91$ for $\gamma_2 = 0.1$. Such odds ratios are not sufficiently away from 1 to enable estimators of φ to be significantly different from 1 for usual sample sizes. Therefore, the negative triple-mutation cannot have much effect on the odds ratio of the collapsed table since γ_2 is usually very small.

When the triple-mutation effect is positive ($\delta_{222} > 1$), $\varphi = (2 + \delta_{222})/3$ for $\gamma_2 = 0.5$, $\varphi = (5 + \delta_{222})/6$ for $\gamma_2 = 0.2$ and $\varphi = (10 + \delta_{222})/11$ for $\gamma_2 = 0.1$. Therefore, the influence of the triple-mutation effect on the odds ratio of the collapsed table is weakened as γ_2 becomes smaller.

When $\delta_{222} = 1$, the conditional association between positions I and II at C_1 and C_2 of position III are both θ_{221}. However, if we collapse position III, the association between positions I and II is

$$\frac{(1+\gamma_2)(1+\gamma_2\theta_{122}\theta_{212})\,\theta_{221}}{(1+\gamma_2\theta_{122})(1+\gamma_2\theta_{212})},$$

which is different from θ_{221}. It will be the same as θ_{221}, if and only if either θ_{122} or θ_{212} equals 1. Therefore, in general, when $\delta_{222} = 1$, conditional associations cannot be interpreted as marginal associations. We cannot collapse any of the three positions to examine the association between the other two positions. In this situation, we should investigate the association between two positions at each level of the third position and thereby address the conditional association.

2.2. Hypotheses

We consider five sets of null hypotheses for multinomially distributed counts n_{ijk}.

(1) H_{01}: Both double mutations and triple mutations are neutral (i.e. $\theta_{122} = \theta_{212} = \theta_{221} = \delta_{222} = 1$).
(2) H_{02}: The triple mutation is neutral (i.e. $\delta_{222} = 1$).
(3) H_{03}: Mutations at two positions are conditionally independent given the other position.
There are three hypotheses of this type, $\theta_{221} = \delta_{222} = 1$, $\theta_{212} = \delta_{222} = 1$ and $\theta_{122} = \delta_{222} = 1$. We previously considered $\theta_{221} = \delta_{222} = 1$ for positions I and II being independent conditionally on position III.
(4) H_{04}: Mutations at two positions are jointly independent of the other position.
There are three hypotheses of this type; $\theta_{212} = \theta_{122} = \delta_{222} = 1$, $\theta_{122} = \theta_{221} = \delta_{222} = 1$ and $\theta_{212} = \theta_{221} = \delta_{222} = 1$. Among them, we consider $\theta_{212} = \theta_{122} = \delta_{222} = 1$, which means that positions I and II are jointly independent of position III.
(5) H_{05}: Double mutations are neutral (i.e. $\theta_{122} = \theta_{212} = \theta_{221} = 1$).

Roy and Kastenbaum (1956) refer to H_{04} as multiple independence:

$$H_0 = p_{ijk} = p_{ij.}\ p_{..k},$$

where $p_{ijk} = E(n_{ijk})/n$ and "." denotes summation over a subscript. They show that it implies

$$p_{i.k} = p_{i..}\ p_{..k} \quad \text{and} \quad p_{.jk} = p_{.j.}\ p_{..k}.$$

However, the latter does not imply H_0 unless the following is also true:

$$p_{ijk} = q_{ij.}\ q_{i.k}\ q_{.jk}$$

for arbitrary functions q's. They term this relationship the hypothesis of "no interaction."

Note that the model which supports H_{05} is not hierarchical since it allows the existence of a triple-mutation effect without any double-mutation effect. Whenever the hierarchical model contains higher-order effects, their lower-order effects are in the model as well. A non-hierarchical model can be problematic since its parameterization depends

UNCORRECTED PROOF

Please cite this article as: Ahn, C., et al., Log-linear models for mutations in the HIV genome, J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.007

on the selection of the reference cell and the choice of reference cell can be arbitrary. As we mentioned previously, we usually have a cell which is dominant over the other cells in the present application, and so it can be used as a meaningful reference cell. Therefore, we can consider a model with only the triple-mutation effect and none of the three double-mutation effects.

2.3. Log-linear models

The $2 \times 2 \times 2$ case can be extended to the more general case with I, J and K possible characters at each position, respectively. Instead of presenting a complicated $I \times J \times K$ table, a log-linear model is a good way to represent its structure.

As discussed in Agresti (2002), the full saturated log-linear model has the following specification:

$$\ln E(n_{ijk}) = m + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}. \tag{1}$$

To take the consensus into consideration, we use the first row, column and depth as the reference cell so that

$$\lambda_i^A = \lambda_j^B = \lambda_k^C = \lambda_{ij}^{AB} = \lambda_{ik}^{AC} = \lambda_{jk}^{BC} = \lambda_{ijk}^{ABC} = 0 \quad \text{when}$$

$$i = 1 \quad \text{or} \quad j = 1 \quad \text{or} \quad k = 1.$$

The relationships between parameters of the log-linear model and those of Table 3 are

$$m = \ln \mu, \quad \lambda_i^A = \ln \alpha_i, \quad \lambda_j^B = \ln \beta_j, \quad \lambda_k^C = \ln \gamma_k, \quad \lambda_{ij}^{AB} = \ln \theta_{ij1}, \quad \lambda_{jk}^{BC} = \ln \theta_{1jk},$$

$$\lambda_{ik}^{AC} = \ln \theta_{i1k} \quad \text{and} \quad \lambda_{ijk}^{ABC} = \ln \delta_{ijk}.$$

We can redefine the five types of null hypotheses, as in Section 2.2, with this new parameterization:

(1) H_{01}: $\lambda_{ij}^{AB} = \lambda_{ik}^{AC} = \lambda_{jk}^{BC} = \lambda_{ijk}^{ABC} = 0$,
(2) H_{02}: $\lambda_{ijk}^{ABC} = 0$,
(3) H_{03}: $\lambda_{ij}^{AB} = \lambda_{ijk}^{ABC} = 0$,
(4) H_{04}: $\lambda_{ik}^{AC} = \lambda_{jk}^{BC} = \lambda_{ijk}^{ABC} = 0$,
(5) H_{05}: $\lambda_{ij}^{AB} = \lambda_{ik}^{AC} = \lambda_{jk}^{BC} = 0$.

2.3.1. Wald statistic

The maximum-likelihood estimators for the parameters in the saturated model (1) are

$$\widehat{m} = \ln n_{111}, \quad \widehat{\lambda}_i^A = \ln \frac{n_{i11}}{n_{111}}, \quad \widehat{\lambda}_j^B = \ln \frac{n_{1j1}}{n_{111}}, \quad \widehat{\lambda}_k^C = \ln \frac{n_{11k}}{n_{111}}, \quad \widehat{\lambda}_{ij}^{AB} = \ln \frac{n_{ij1} n_{111}}{n_{i11} n_{1j1}},$$

$$\widehat{\lambda}_{jk}^{BC} = \ln \frac{n_{1jk} n_{111}}{n_{1j1} n_{11k}}, \quad \widehat{\lambda}_{ik}^{AC} = \ln \frac{n_{i1k} n_{111}}{n_{i11} n_{11k}} \quad \text{and} \quad \widehat{\lambda}_{ijk}^{ABC} = \ln \frac{n_{ijk} n_{i11} n_{1j1} n_{11k}}{n_{111} n_{ij1} n_{1jk} n_{i1k}}.$$

To test H_{01}, λ_{ij}^{AB}, λ_{ik}^{AC}, λ_{jk}^{BC} and λ_{ijk}^{ABC} should be tested simultaneously. The Wald statistic for H_{01}, Q_W, is

$$(\widehat{\lambda}_{ij}^{AB}, \widehat{\lambda}_{ik}^{AC}, \widehat{\lambda}_{jk}^{BC}, \widehat{\lambda}_{ijk}^{ABC}) \{\widehat{Var}(\widehat{\lambda}_{ij}^{AB}, \widehat{\lambda}_{ik}^{AC}, \widehat{\lambda}_{jk}^{BC}, \widehat{\lambda}_{ijk}^{ABC})\}^{-1} (\widehat{\lambda}_{ij}^{AB}, \widehat{\lambda}_{ik}^{AC}, \widehat{\lambda}_{jk}^{BC}, \widehat{\lambda}_{ijk}^{ABC})^{\tau}.$$

We can construct the test statistics for the other hypotheses in the same way. This type of Wald statistic has reasonable statistical behavior with respect to the saturated model in terms of an approximately chi-square distribution (with degrees of freedom equal to the dimension of the hypothesis) when all $n_{ijk} \geqslant 5$. For situations where some n_{ijk} are 2, 3, or 4 and most $\geqslant 5$, the deviance test in the next section can have somewhat better behavior for hypotheses pertaining

UNCORRECTED PROOF

Please cite this article as: Ahn, C., et al., Log-linear models for mutations in the HIV genome, J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.007

to the saturated model, but exact methods become necessary for accurate assessment when many n_{ijk} are < 5 or some are < 2.

2.3.2. Deviance test

For log-linear models, the likelihood-ratio statistic is the more usual choice for testing the hypotheses in this section. The likelihood-ratio statistic of interest is -2 times the logarithm of the likelihood ratio for the null model and the full saturated model. Therefore, the test statistic is the deviance, Q_{L} say, since the full model is always the saturated model in this case. The deviance has an approximately chi-square distribution with degrees of freedom being the difference between the number of parameters in the full model and that in the null model.

The saturated model is

$$\log E(n_{ijk}) = m + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ},$$

and the reduced models are

(1) for H_{01}

$$\log E(n_{ijk}) = m + \lambda_i^X + \lambda_j^Y + \lambda_k^Z,$$

(2) for H_{02}

$$\log E(n_{ijk}) = m + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ},$$

(3) for H_{03}

$$\log E(n_{ijk}) = m + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ},$$

(4) for H_{04}

$$\log E(n_{ijk}) = m + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY},$$

(5) for H_{05}

$$\log E(n_{ijk}) = m + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ijk}^{XYZ}.$$

The maximum-likelihood estimators for the parameters required to test H_{01}, H_{03} and H_{04} have explicit formulae, while those for H_{02} and H_{05} need an iterative method such as the Newton–Raphson method (Agresti, 2002).

3. Examples

We consider 492 non-syncytium inducing (NSI) HIV-1 sequences from clade B spanning the 35 amino acids of the V3 loop of the envelope protein, the most variable part of the HIV genome and the target of vaccine development. Ahn (2004) showed that mutations at positions 11, 14, 16, 18, 19 and 20 have strong associations.

If we consider protein sequences, there are 20 possible amino acids at each position. At most positions, we observe about 10 amino acids. Here, all analyses are performed with positions as dichotomous variates (consensus vs. non-consensus) for simplicity and feasibility of model fitting (although consideration could be comparably given to dichotomous variates for the consensus vs. the most likely non-consensus outcome or to categorical variates for consensus vs. one or more non-consensus outcomes with enough sample size for their assessment to be informative).

We examine four triplets of the six positions as illustrative examples in Section 3.1 for tests of the hypotheses in Section 2. We do not address the first hypothesis since a completely independent triplet is not present for the six positions under consideration, mainly because these six positions are of interest through their strong associations with one another. Since the illustrative triplets in these analyses are from collapsings of the other three of six positions, their results should be viewed with some caution relative to the issues in Section 2. Accordingly, in Section 3.2, we consider a log-linear model including all six positions. It helps us to understand all the complex interrelationships which apply.

UNCORRECTED PROOF

Please cite this article as: Ahn, C., et al., Log-linear models for mutations in the HIV genome, J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.007

3.1. Three-way tables

We investigate three-way tables for the six positions, since the sample size is not large enough to be informative for higher-order associations. In this regard, only 45 of the $2^6 = 64$ possible degrees of freedom for the respective levels of association are available for assessment. Moreover, in Section 3.2, reasonable goodness of fit is provided for the six-way table by a log-linear model corresponding to three three-way tables and four two-way tables which are non-redundant. Results from four three-way tables are presented as examples to illustrate various types of association with respect to the five hypotheses in Section 2.2.

For (11,14,16), the likelihood-ratio statistic supports a version of H_{04}, $\theta_{221} = \theta_{212} = \delta_{222} = 1$ ($Q_L(\text{df} = 3) = 2.78$, $p = 0.43$). This result supports positions 14 and 16 being jointly independent of position 11.

For (14,18,20), the likelihood-ratio statistic supports H_{02}, $\delta_{222} = 1$ ($Q_L(\text{df} = 1) = 0.79$, $p = 0.37$). This result indicates no triple-mutation effect in the presence of all double-mutation effects.

For (14,16,19), the likelihood-ratio statistic supports a version of H_{03}, $\theta_{122} = \delta_{222} = 1$ ($Q_L(\text{df} = 2) = 0.36$, $p = 0.84$). This result is compatible with the plausibility of the hypothesis that positions 16 and 19 are conditionally independent for each category of position 14.

For (14,16,18), the likelihood-ratio statistic for H_{02}, $\delta_{222} = 1$ ($Q_L(\text{df} = 1) = 5.87$, $p = 0.015$) suggests the presence of a triple-mutation effect. However, the frequency table for these three positions has $n_{122} = 1$ and $n_{222} = 0$ which are too small to support an approximately chi-square distribution for Q_L. With the use, of exact logistic regression (as described in Stokes et al., 2000, Chapter 8), an exact 0.95 confidence interval for δ_{222} is $(0.433, \infty)$ with 0.112 as the corresponding p-value. Thus, the exact analysis suggests compatibility of (14,16,18) with H_{02} (although interpretation here needs some caution because of potentially low power for the assessment of H_{02}). With the use of a log-linear model that only includes double-mutation effects (by removal of δ_{222}), exact methods also have some relevance to the assessment of θ_{122} since $n_{\cdot 22} = 1$. In this case, the exact 0.95 confidence interval is $(8.90, > 1000)$ from exact logistic regression whereas its approximate counterpart from the usual maximum-likelihood methods is (7.46,409.33). Thus, although the exact result provides a more accurate assessment, its approximate counterpart is comparable for most practical purposes. Nevertheless, exact methods should be viewed as preferable when their application is feasible.

3.2. Considering all positions

Three-way tables might be over-simplifications for understanding the overall associations in the mutually correlated positions 11, 14, 16, 18, 19 and 20. Therefore, we examine all of these six positions simultaneously. However, due to limitations of sample size relative to available degrees of freedom, we do not consider interactions of order higher than three-way interactions (since only 45 of the possible 64 degrees of freedom are available for assessment).

We fit a log-linear model with all two-way interactions via maximum-likelihood methods for Poisson regression as described in Chapters 12, 16 and 17 of Stokes et al. (2000). Then, we start to add the three-way interactions one by one if their Wald statistic p-value is less than 0.05 (except the three-way interaction corresponding to (14,16,18) for which exclusion was based on the exact assessment discussed in Section 3.1). Then, we eliminate the two-way interactions one by one if their p-value is > 0.05 and they are not part of a retained three-way interaction with p-value < 0.05. Ultimately, a hierarchical log-linear model is reached with inclusion of all two- and three-way interactions with p-value < 0.05 (or nearly so) as well as any two-way interaction nested within included three-way interactions. The resulting model has 20 degrees of freedom for two-way and three-way associations, and its goodness of fit is supported by the likelihood-ratio (deviance) test with $p = 0.25$ for the 24 degrees of freedom for the higher-order interactions which are excluded. Some caution is needed for interpreting this assessment of goodness of fit because the sample size is not really large enough to support the test statistic having either an approximately chi-square distribution or adequate power, but the model is still useful in the sense of including three-way and two-way interactions that were not removed by the stepwise process.

In Table 5, we find three-way interactions for (11,18,19), (14,18,19), and (14,18,20) with p-values < 0.05. Among the two-way interactions which are included in the noted three-way interactions, all have p-values < 0.05 except those for (14,20) and (18,20). Additional two-way interactions with p-values < 0.05 are (14,16), (16,18), (16,20), and (19,20).

Since the coefficients for the (11,18,19) and (14,18,19) three-way interactions in the model are negative, the corresponding triple mutations have frequencies which are lower than what would be expected from the frequencies of double mutations corresponding to the included two-way interactions. Conversely, the coefficient for (14,18,20) is

UNCORRECTED PROOF

Please cite this article as: Ahn, C., et al., Log-linear models for mutations in the HIV genome, J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.007

Table 5
Estimates (Est.), standard errors (SE), deviance test statistics (Q_L), and p-values for a hierarchical log-linear model for a cluster of six positions

Positions	Est.	SE	Q_L	p-value
11	−1.17	0.13	90.69	< 0.001
14	−1.49	0.16	102.51	< 0.001
16	−3.32	0.34	238.79	< 0.001
18	−1.47	0.18	83.97	< 0.001
19	−3.42	0.34	223.52	< 0.001
20	−1.52	0.17	104.83	< 0.001
(11,18)	0.54	0.24	5.08	0.024
(11,19)	0.78	0.34	5.01	0.025
(14,16)	2.89	0.37	83.16	< 0.001
(14,18)	1.07	0.28	14.80	< 0.001
(14,19)	1.68	0.36	23.32	< 0.001
(14,20)	0.01	0.32	0.00	0.97
(16,18)	−4.18	1.03	57.84	< 0.001
(16,20)	0.73	0.35	4.32	0.038
(18,19)	1.67	0.47	12.13	< 0.001
(18,20)	0.16	0.33	0.24	0.62
(19,20)	1.11	0.29	14.52	< 0.001
(11,18,19)	−1.20	0.62	3.93	0.047
(14,18,19)	−2.12	0.60	12.93	< 0.001
(14,18,20)	1.01	0.50	4.28	0.039
Goodness-of-fit likelihood-ratio test	(df = 24)		28.37	0.25

positive, and so the frequency for its triple mutation is higher than expected. Also, since there are positive signs for all of the coefficients for the two-way interactions within the three-way interactions of the model, double mutations for the corresponding pairs of positions at the consensus of the third position have frequencies which are higher than expected, although this tendency is essentially null for (14,20) and (18,20). The tendency for higher than expected frequencies also applies to double mutations for (14,16), (16,20), and (19,20) for all combinations of categories for the other positions, but the opposite tendency for lower than expected frequencies applies to (16,18) because of the negative sign for its coefficient. Otherwise, (14,18,20) is nearly compatible with H_{05} in the sense that its triple mutation has $p < 0.05$ and two of the double mutations within it are essentially null at the consensus with $p > 0.50$.

For the analysis in Table 5, the two-way interactions within the three-way interactions of the model correspond to interpretable parameters from reference-cell coding as in Table 3. However, for counterparts to such two-way interactions with averaging over the third position, the likelihood-ratio test statistics and p-values are $Q_L = 0.04$ with $p = 0.85$ for (11, 18), $Q_L = 0.35$ with $p = 0.55$ for (11, 19), $Q_L = 2.66$ with $p = 0.10$ for (14, 18), $Q_L = 4.17$ with $p = 0.041$ for (14, 19), $Q_L = 4.43$ with $p = 0.035$ for (14, 20), $Q_L = 0.00$ with $p = 0.98$ for (18, 19), and $Q_L = 7.07$ with $p = 0.008$ for (18, 20). Since the coefficients for the three-way interactions for (11, 18, 19) and (14, 18, 19) are negative and those for the two-way interactions within them (for pairs of positions at the consensus of the third) are positive, such results for averaging counterparts to two-way interactions tend to be uninterpretable from the conflict of the positive two-way associations for pairs of positions at the consensus vs. negative associations at the corresponding non-consensus, particularly for (11, 18), (11, 19), (14, 18), and (18, 19) where no two-way association is suggested by the results which involve averaging. Relative to the (14, 18, 20) three-way interaction for which the coefficient is positive, there tends to be larger positive two-way association for pairs of positions at the non-consensus at the third position than at the consensus, particularly for (14, 20) and (18, 20). As stated previously, careful consideration is needed for distinguishing between the results of the tests in Table 5 for the interpretable parameters from reference-cell coding for two-way interactions within a three-way interaction in the model vs. their often uninterpretable counterparts from averaging across the third position (of the three-way interaction which includes them). Clearly, to understand the effect of mutations on viral survival, one must have assessments that are beyond pairwise considerations.

Likelihood-ratio (or deviance) test statistics were used to assess the interactions in Table 5, but the Wald statistic could alternatively have been used. For the non-redundant three-way and two-way interactions which correspond

UNCORRECTED PROOF

Please cite this article as: Ahn, C., et al., Log-linear models for mutations in the HIV genome, J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.007

to the model, the respective p-values from Wald statistics are < 0.001 for (14, 16), < 0.001 for (16, 18), 0.037 for (16, 20), < 0.001 for (19, 20), 0.053 for (11, 18, 19), < 0.001 for (14, 18, 19) and 0.041 for (14, 18, 20), and so have comparable interpretations to their deviance counterparts. This comparability is expected in view of the asymptotic equivalence of the Wald statistic and the deviance for assessments pertaining to a reduced model like that in Table 5 when the corresponding non-redundant three-way and two-way tables corresponding to that model have sufficiently large cell counts (see Imrey et al., 1981, 1982). The full six-way table is given in the Appendix with the non-redundant three-way and two-way marginal tables corresponding to the model. The cell counts in these respective marginal tables are considered large enough to support the comparable use of the Wald statistic or the deviance as described for this example, although the presence of a cell count of 1 in the (16, 18) marginal table enables the deviance to be somewhat preferable.

4. Discussion

The methodology presented in this paper addresses a multidimensional contingency table which can have complicated interpretations of its parameters. Such complexity is a reason for collapsing some dimensions into a simpler table. However, as we discussed, we should be very careful about collapsing some dimensions, since the parametric structure does not remain the same. The parameters in the collapsed table only retain the same meaning as in the full table when collapsibility conditions are satisfied as in Section 2.1.

In addition, a higher-dimensional table tends to possess many empty cells. Therefore, when we perform data analysis, we might consider all variates as dichotomous by combining all non-consensus characters to avoid overly small cell counts. This process may mask possible associations if the associations in the combined cells have conflicting patterns which cancel each other. An exact method can be a reasonable approach for this matter. Even though evaluating all possible reference tables to calculate an exact p-value is computationally demanding, the improving modern technology will make it easier. Gibbs sampling is a possibility (Forster et al., 1996). However, in most situations, we have no choice but to collapse some variates or combine some characters due to computational difficulty and model feasibility.

The global variation in HIV has been broken down into groups (called *clades*). When we investigate the association over different clades, the clade identity should be considered as a stratification variate. A stratified two-way table can be also analyzed as a special case of a three-way table. Therefore, we can still apply some of the methods discussed here in the presence of a stratification variate.

5. Uncited reference

Bishop (1971).

Appendix A.

The full six-way table is given in Tables A1–A6.

Table A1
Position 11 × 18 × 19 vs. Position 14 × 16 × 20

		14 × 16 × 20							
		111	112	121	122	211	212	221	222
11 × 18 × 19	111	135	27	2	4	27	6	24	13
	112	3	3	0	0	7	6	2	4
	121	31	7	1	0	21	15	0	0
	122	6	4	0	0	2	5	0	0
	211	45	10	2	0	9	2	5	1
	212	4	2	0	1	4	2	3	1
	221	16	5	0	0	11	8	0	0
	222	1	2	0	0	1	2	0	0

UNCORRECTED PROOF

Please cite this article as: Ahn, C., et al., Log-linear models for mutations in the HIV genome, J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.007

Table A2
Position 11× Position 18× Position 19

Position 11		1		2	
Position 18		1	2	1	2
Position 19	1	238	25	74	17
	2	75	17	40	6

Table A3
Position 14× Position 18× Position 20

Position 14		1		2	
Position 18		1	2	1	2
Position 20	1	191	47	81	35
	2	55	18	35	30

Table A4
Position 14× Position 18× Position 19

Position 14		1		2	
Position 18		1	2	1	2
Position 19	1	225	13	87	29
	2	60	13	55	10

Table A5
Position 16× Position 19× Position 20

Position 16		1		2	
Position 19		1	2	1	2
Position 20	1	295	80	34	18
	2	28	26	5	6

Table A6
Position 14× Position 16× Position 18

Position 14		1		2	
Position 16		1	2	1	2
Position 18	1	229	72	63	65
	2	9	1	53	0

UNCORRECTED PROOF

References

Agresti, A., 2002. Categorical Data Analysis. second ed.. Wiley, New York.

Ahn, C., 2004. Detecting linked changed in fast evolving genome. Ph.D. Thesis, Department of Biostatistics, University of North Carolina at Chapel Hill, NC.

Bhapkar, V., Koch, G., 1968. On the hypotheses of 'no interaction' in contingency tables. Biometrics 24, 567–594.

Bishop, Y., 1971. Effects of collapsing multidimensional contingency tables. Biometrics 27.

Please cite this article as: Ahn, C., et al., Log-linear models for mutations in the HIV genome, J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.007

Forster, J., McDonald, J., Smith, P., 1996. Monte Carlo exact conditional tests for log-linear and logistic models. J. Roy. Statist. Soc. Ser. B 58, 445–453.

Imrey, P., 2000. Poisson regression, logistic regression, and loglinear models for random counts. In: Tinsley, H., Brown, S. (Eds.), Handbook of Applied Multivariate Statistics and Mathematical Modeling. Academic Press, San Diego, CA, pp. 391–437.

Imrey, P., Koch, G., 2005. Categorical data analysis. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of Biostatistics, second ed., vol. 1. Wiley, New York, pp. 682–707.

Imrey, P., Koch, G., Stokes, M., Darroch, J., Freeman, D., Tolley, H., 1981. Categorical data analysis: some reflections on the log linear model and logistic regression, part i. Internat. Statist. Rev. 49, 265–283.

Imrey, P., Koch, G., Stokes, M., Darroch, J., Freeman, D., Tolley, H., 1982. Categorical data analysis: some reflections on the log linear model and logistic regression, part ii. Intetnat. Statist. Rev. 50, 35–64.

Imrey, P., Koch, G., Preisser, J., 1996. The evolution of categorical data modeling: a biometric perspective. In: David, H., Armitage, P. (Eds.), Advances in Biometry. Wiley, New York, pp. 89–114.

Roy, S., 1957. Some Aspects of Multivariate Analysis. Wiley, New York.

Roy, S., Bhapkar, V., 1960. Some nonparametric analogues of 'normal' ANOVA, MANOVA, and of studies in 'normal' association. In: Olkin, I., Ghurye, S., Hoeffding, W., Madow, W., Mann, H. (Eds.), Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford University Press, Stanford, CA, pp. 371–387.

Roy, S., Kastenbaum, M., 1956. On the hypothesis of no "interaction" in a multi-way contingency table. Ann. Math. Statist. 27, 749–757.

Roy, S., Mitra, S., 1956. An introduction to some non-parametric generalizations of analysis of variance and multivariate analysis. Biometrika 43, 361–376.

Stokes, M., Davis, C., Koch, G., 2000. Categorical Data Analysis Using the SAS System. second ed.. SAS Institute, Cary, NC.

UNCORRECTED PROOF

Please cite this article as: Ahn, C., et al., Log-linear models for mutations in the HIV genome, J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.007